# Contents

## Preface

## I Foundations  1

### 1 Introduction: Fundamental Concepts and the Human Genome  3

#### Objectives  3

1.1 Introduction  3

#### Motivation and aim of this book  3

1.2 Overview of topics covered in this book  6

1.3 What are DNA, the genome, a gene, and a chromosome?  8

#### Mendel's laws, sexual reproduction, and genetic recombination  9

1.4 From genes to protein and the central dogma of molecular biology  15

#### From genes to protein: Genes, amino acids, nucleotides, and proteins  15

1.5 Homozygous and heterozygous alleles, dominant and recessive traits  20

1.6 Heritability  22

#### Defining heritability: Broad- and narrow-sense heritability  22

1.7 Conclusion  28

#### Exercises  28

#### Further reading and resources  29

#### References  30

## 2 A Statistical Primer for Genetic Data Analysis  33

#### Objectives  33

2.1 Introduction  33

2.2 Basic statistical concepts  34
## Table of Contents

### 2.2.1 Mean, standard deviation, and variance 34
### 2.2.2 Covariance and the variance-covariance matrix 36
### 2.3 Statistical models 38
#### 2.3.1 Regression models 38
#### 2.3.2 The null and alternative hypothesis and significance thresholds 39
### 2.4 Correlation, causation, and multivariate causal models 40
#### 2.4.1 Correlation versus causation 40
#### 2.4.2 Multivariate causal models 42
### 2.5 Fixed-effects models, random-effects models, and mixed models 47
### 2.6 Replication of results and overfitting 48
### 2.7 Conclusion 49

### Exercises 50

### Further reading 52

### Software for mixed-model analyses 52

### Appendix 52

### References 54

### 3 A Primer in Human Evolution 55

#### Objectives 55

#### 3.1 Introduction 55

#### 3.2 Human dispersal out of Africa 56

#### 3.3 Population structure and stratification 58
##### 3.3.1 Population structure, genetic admixture, and Principal Component Analysis (PCA) 58
##### 3.3.2 Common misnomers of population structure: Ancestry is not race 59
##### 3.3.3 Genetic scores cannot be transferred across ancestry groups 59
##### 3.3.4 How genes mirror geography 61

#### 3.4 Human evolution, selection, and adaptation 63
##### 3.4.1 Evolution, fitness, and natural selection 63
##### 3.4.2 Genetic drift 68

#### 3.5 The Hardy–Weinberg equilibrium 69
##### 3.5.1 Assumptions of the HWE 69
##### 3.5.2 Understanding the notation of the HWE 70

#### 3.6 Linkage disequilibrium and haplotype blocks 71

#### 3.7 Conclusion 73

### Exercises 73

### Further reading and resources 74

### References 74

### 4 Genome-Wide Association Studies 77

#### Objectives 77

#### 4.1 Introduction and background 77

#### 4.2 GWAS research design and meta-analysis 79
##### 4.2.1 GWAS research design 79
##### 4.2.2 Data analysis plan 81
##### 4.2.3 Meta-analysis 82

### References 82
4.3 Statistical inference, methods, and heterogeneity  83
  4.3.1 Nature of the phenotype  83
  4.3.2 P-values and Z-scores  83
  4.3.3 Correcting for multiple testing in a GWAS  84
  4.3.4 Manhattan plots  85
  4.3.5 Evaluating dichotomous versus quantitative traits  87
  4.3.6 Fixed-effects versus random-effects models  88
  4.3.7 Weighting, false discovery rate (FDR), and imputation  89
  4.3.8 Sources of heterogeneity  89
4.4 Quality control (QC) of genetic data  90
4.5 The NHGRI-EBI GWAS Catalog  91
  4.5.1 What is the NHGRI-EBI GWAS Catalog?  91
  4.5.2 A brief history of the GWAS  91
  4.5.3 Lack of diversity in GWASs  93
4.6 Conclusion and future directions  97

Exercises  98
Further reading  98
References  99

5 Introduction to Polygenic Scores and Genetic Architecture  101

Objectives  101
5.1 Introduction  101
  5.1.1 What is a polygenic score?  105
  5.1.2 The origins of polygenic scores  105
5.2 Construction of polygenic scores  107
  5.2.1 Large sample sizes required in GWAS discovery  108
  5.2.2 Selection of SNPs to include  108
5.3 Validation and prediction of polygenic scores  108
  5.3.1 Independent target sample  109
  5.3.2 Similar ancestry in target sample  110
  5.3.3 Relatedness, population stratification, and differential bias  110
  5.3.4 Variance explained only by common genetic markers missing rare variants  111
  5.3.5 Missing and hidden heritability in prediction of phenotypes from genetic markers (SNPs)  111
  5.3.6 Trade-off between prediction and understanding biological mechanisms  112
5.4 Shared genetic architecture of phenotypes  113
  5.4.1 Predicting other phenotypes  113
  5.4.2 Phenotypic and genetic correlation  114
  5.4.3 Pleiotropy  115
  5.4.4 Multitrait analysis  119
5.5 Causal modeling with polygenic scores  119
  5.5.1 Genetic confounding  119
  5.5.2 Mendelian Randomization  120
  5.5.3 Controlling for confounders  120
  5.5.4 Gene-environment interaction and heterogeneity  122
6 Gene-Environment Interplay  129
Objectives  129
6.1 Introduction: What is gene-environment (G×E) interplay?  129
6.2 Defining the environment in G×E research  130
  6.2.1 Nature and scope of E: Multilevel, multidomain, and multitemporal  131
  6.2.2 Interdependence of environmental risk factors  132
6.3 A brief history of G×E research  133
  6.3.1 Classic approaches  133
  6.3.2 Candidate gene cG×E approaches  134
  6.3.3 Genome-wide polygenic score G×E approaches  135
6.4 Conceptual G×E models  136
  6.4.1 Diathesis-stress, vulnerability, or contextual triggering model  136
  6.4.2 Bioecological or social compensation model  137
  6.4.3 Differential susceptibility model  139
  6.4.4 Social control or social push model  140
  6.4.5 Research designs to study G×E  140
6.5 Gene-environment correlation (rGE)  143
  6.5.1 Passive gene-environment correlation (rGE)  144
  6.5.2 Evocative (or reactive) rGE  145
  6.5.3 Active rGE  145
  6.5.4 Why are models of rGE important?  145
  6.5.5 Research designs to study rGE  146
6.6 Conclusion and future directions  146
  6.6.1 Why haven’t many G×E s been identified?  146
Exercises  147
Further reading  147
References  147

II Working with Genetic Data  151

7 Genetic Data and Analytical Challenges  153
Objectives  153
7.1 Introduction  153
7.2 Genotyping and sequencing array  154
  7.2.1 Genotyping and sequencing technologies  154
  7.2.2 Linkage disequilibrium and imputation  155
  7.2.3 Limitations of genotyping arrays and next-generation sequencing  158
  7.2.4 Drop in costs per genome  159
7.3 Overview of human genetic data for analysis  160
  7.3.1 Prominently used genetic data  161
  7.3.2 Sources that archive and distribute data  163
  7.3.3 Obtaining GWAS summary statistics  164
7.4 Different formats in genomics data 165
  7.4.1 Genomics data is big data 165
  7.4.2 PLINK software and genotype formats 166
  7.4.3 PLINK binary files 170
7.5 Genetic formats for imputed data 171
  7.5.1 PLINK 2.0 171
  7.5.2 Oxford file formats 172
  7.5.3 The variant call format (VCF) 174
7.6 Data used in this book 175
7.7 Data transfer, storage, size, and computing power 176
  7.7.1 Data storage 176
  7.7.2 Data sharing, transfer across borders, and cloud storage 177
  7.7.3 Size of data and computational power 178
7.8 Conclusion 179
Exercises 179
Further reading and resources 179
References 180

8 Working with Genetic Data, Part I: Data Management, Descriptive Statistics,
and Quality Control 183
Objectives 183
8.1 Introduction: Working with genetic data 183
8.2 Getting started with PLINK 184
  8.2.1 The command line 184
  8.2.2 Calling PLINK and the PLINK command line 186
  8.2.3 Running scripts in terminal 188
  8.2.4 Opening PLINK files 189
  8.2.5 Recode binary files to create new readable dataset with .ped
      and .map files 189
  8.2.6 Import data from other formats 191
8.3 Data management 193
  8.3.1 Select individuals and markers 193
  8.3.2 Merge different genetic files and attaching a phenotype 196
8.4 Descriptive statistics 199
  8.4.1 Allele frequency 199
  8.4.2 Missing values 200
8.5 Quality control of genetic data 202
  8.5.1 Per-individual QC 203
  8.5.2 Per-marker QC 206
  8.5.3 Genome-wide association meta-analysis QC 209
8.6 Conclusion 211
Exercises 214
Further reading and resources 214
References 214
12 Applying Genome-Wide Association Results 315
Objectives 315
12.1 Introduction 315
12.2 Plotting association results 316
  12.2.1 Manhattan plots 316
  12.1.2 Regional association plots 320
  12.1.3 Quantile-Quantile plots and the \( \lambda \) statistic 320
12.2 Estimating heritability from summary statistics 324
12.3 Estimating genetic correlations from summary statistics 328
12.4 MTAG: Multi-Trait Analysis of Genome-wide association summary statistics 333
12.5 Conclusion 336
Exercises 336
Further reading and resources 336
References 337

13 Mendelian Randomization and Instrumental Variables 339
Objectives 339
13.1 Introduction 339
13.2 Randomized control trials and causality 341
13.3 Mendelian Randomization 341
13.4 Instrumental variables and Mendelian Randomization 343
  13.4.1 The IV model in an MR framework 343
  13.4.2 Violation of statistical assumptions of the IV approach 347
13.5 Extensions of standard MR 349
  13.5.1 Using multiple markers as independent instruments 351
  13.5.2 Using polygenic scores as IVs 351
  13.5.3 Bidirectional MR analyses 352
13.6 Applications of MR 352
  13.6.1 Consequences of alcohol consumption 352
  13.6.2 Body mass index and mortality 353
  13.6.3 Causes of dementia and Alzheimer’s disease 354
13.7 Conclusion 355
Exercises 355
Further reading 356
References 356

14 Ethical Issues in Genomics Research 359
Objectives 359
14.1 Introduction 359
14.2 Genetics is not destiny: Genetic determinism 361
  14.2.1 Variation in traits and ability to use individual PGSs as predictors 361
  14.2.2 Heritability and missing heritability 362
14.3 Clinical use of PGSs 363
  14.3.1 Genetics and family history 363
  14.3.2 Genetic scores for screening, intervention, and life planning 364
14.3.3 Pharmacogenetics  365
14.3.4 Public understanding of genetic information and information risks  366
14.4 Lack of diversity in genomics  367
14.4.1 Lack of diversity in GWASs  367
14.4.2 European ancestry bias related to PGS construction  367
14.5 Privacy, consent, legal issues, insurance, and General Data Protection Regulation  367
14.5.1 Privacy in the age of public genetics: Solving crimes and finding people  367
14.5.2 The changing nature of informed consent in genomic research  368
14.5.3 Insurance and genetics  369
14.5.4 GDPR and genetics  370
14.6 Conclusion and future directions  372
Further reading and resources  373
References  373

15 Conclusions and Future Directions  377
15.1 Summary and reflection  377
15.2 Future directions  377
References  380

Appendix 1: Software Used in This Book  381
A1.1 Introduction  381
A1.2 RStudio and R  381
A1.3 PLINK  382
A1.4 GCTA  382
A1.5 PRSice  382
A1.6 Python  383
A1.6.1 How to switch from Python 3 to Python 2  384
A1.6.2 Installing packages in Python  385
A1.7 Git  385
A1.8 LDpred  386
A1.9 LDSC  386
A1.10 MTAG  387
A1.11 Using Windows for this book  388
References  388

Appendix 2: Data Used in This Book  389
A2.1 Introduction  389
A2.2 Description of simulated data  389
A2.3 Health and Retirement Study  391
A2.4 Data used by chapter  395
References  397

Glossary  399
Notes  405
Index  409